**大數據分析與實作\_期末作業**

指導教授：龔千芬 老師

組別：第十一組

組員：

C108156122 王薈宣

C108156135 鄭繼威

C108156140 葉桔良

C108156144 吳岳峰

1. 某海產業者想知道雌雄鮑魚重量有無差異？
2. Data Source：<https://archive.ics.uci.edu/ml/datasets/Abalone>

資料描述：性別分為F（公），M（母），I（嬰兒無法分辨）

重量指的是一顆鮑魚帶殼的重量單位為公斤

1. 分析方法：獨立樣本T檢定

為何採用：X為性別（名目尺度），Y為鮑魚含殼重量（連續尺度）

1. 分析結果：

程式碼（Abalone\_ind.py）：

from scipy import stats

import pandas as pd

df1 = pd.read\_excel('Abalone.xlsx')

# print(df1)

F = []

M = []

I = []

for i in range(len(df1['Gender'])):

    if df1['Gender'][i] == "F":

        F.append(df1['Whole\_weight'][i])

    elif df1['Gender'][i] == "M":

        M.append(df1['Whole\_weight'][i])

    else:

        I.append(df1['Whole\_weight'][i])

lev = stats.levene(F, M, center='mean')

print('lev：', lev)

if lev[1] > 0.05:

    ay6t = stats.ttest\_ind(F, M)

    print('變異數同質：', ay6t)

else:

    ay6tf = stats.ttest\_ind(F, M, equal\_var=False)

    print('變異數異質：', ay6tf)

統計假設：

H0：μ雌 = μ雄

H1：μ雌 != μ雄

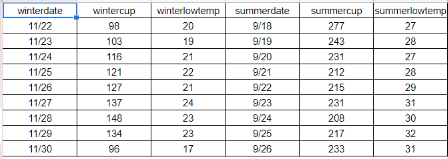
1. 結論：



P = 0.0011550413625208316 < α = 0.05

拒絕H0 =>

在α = 0.05，雌雄鮑魚重量無差異的說法是不可以支持的。

* 1. 某飲料業者想知道夏季與冬季時在飲料銷售（杯數）有無差異？
  2. Data Source：<https://reurl.cc/bkkGRr>

資料描述：這裡面的資料是分別取飲料店夏天與冬天某幾天晚上販賣的杯數。

* 1. 分析方法：獨立樣本Ｔ檢定

為何採用：X為季節（名目尺度），Y為飲料杯數（連續尺度）

* 1. 分析結果：

一張含有 文字 的圖片

自動產生的描述程式碼（drinks.py）：

統計假設：

H0：μ冬季 = μ夏季

H1：μ冬季 != μ夏季

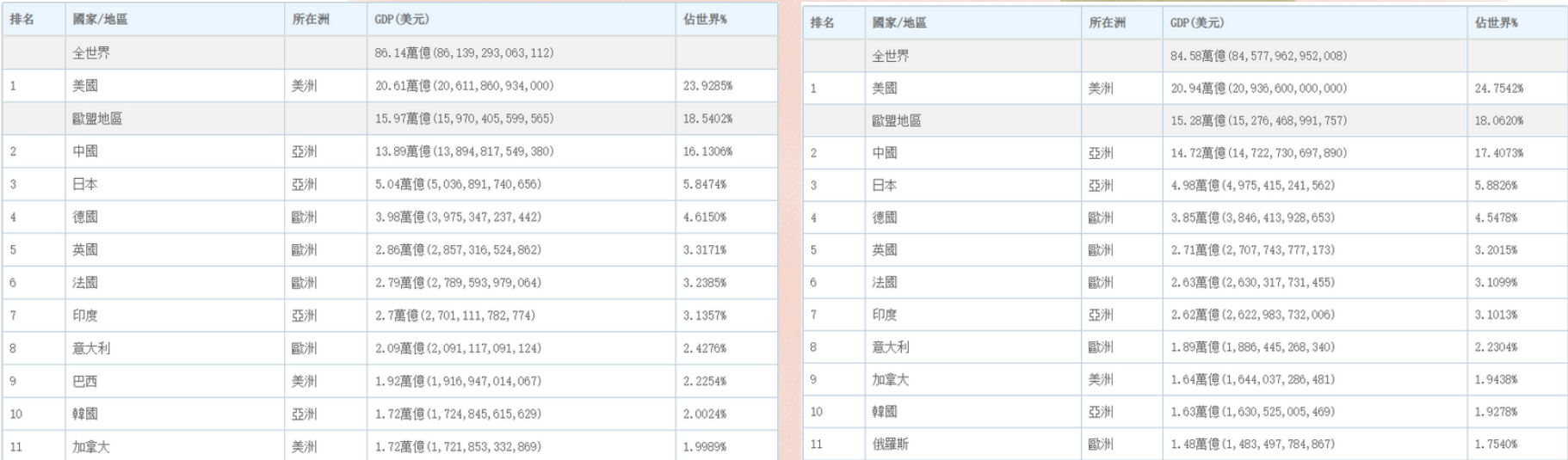
* 1. 結論：

P = 2.795986357901562e-09 < α = 0.05

拒絕H0 =>

在α=0.05，兩季的飲料銷售數量無差異的說法是不可以支持的。

* 1. 在發生疫情前與發生疫情後各國GDP是否有差異？
  2. Data Source：[https://reurl.cc/Q6K7A2](https://reurl.cc/Q6K7A2" \t "_blank) & <https://reurl.cc/82xyMM>



資料描述：每個國家所對應疫情前後的GDP

* 1. 分析方法：獨立樣本T檢定

為何採用：X為年份（名目尺度），Y為各國GDP（連續尺度）

* 1. 分析結果：

程式碼（GDP\_ind.py）：

from scipy import stats

import pandas as pd

df1 = pd.read\_excel('GDP.xlsx')

GDP2018 = df1['2018GDP(美元)']

GDP2020 = df1['2020GDP(美元)']

# print(before)

lev = stats.levene(GDP2018, GDP2020, center='mean')

print('lev：', lev)

if lev[1] > 0.05:

    ay6t = stats.ttest\_ind(GDP2018, GDP2020)

    print('變異數同質：', ay6t)

else:

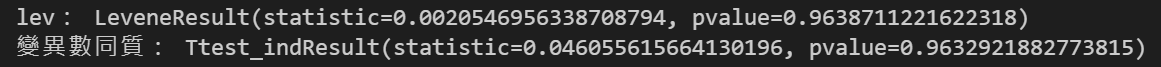
    ay6tf = stats.ttest\_ind(GDP2018, GDP2020, equal\_var=False)

    print('變異數異質：', ay6tf)

統計假設：

H0：各國GDP無差異

H1：各國GDP有差異

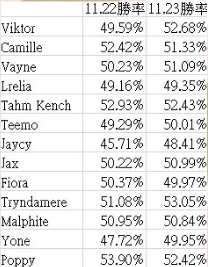
* 1. 結論：

P = 0.9632921882773815 > α = 0.05

接受H0 =>

在α = 0.05，疫情前後GDP無差異的說法是可以支持的。

* 1. 某遊戲為了平衡，在角色技能做了一些調整，請問在更新前與更新後角色勝率是否有差異？
  2. Data Source：<https://tw.op.gg/champion/statistics>



資料描述：分為11.22及11.23版本的角色勝率

* 1. 分析方法：獨立樣本T檢定

為何採用：X為版本（名目尺度），Y為各英雄勝率（連續尺度）

* 1. 分析結果：

程式碼（Hero\_ind.py）：

from scipy import stats

import pandas as pd

df1 = pd.read\_excel('hero.xlsx')

win1122 = df1['11.22勝率']

win1123 = df1['11.23勝率']

# print(before)

lev = stats.levene(win1122, win1123, center='mean')

print('lev：', lev)

if lev[1] > 0.05:

    ay6t = stats.ttest\_ind(win1122, win1123)

    print('變異數同質：', ay6t)

else:

    ay6tf = stats.ttest\_ind(win1122, win1123, equal\_var=False)

    print('變異數異質：', ay6tf)

統計假設：

H0：μ更新前 = μ更新後

H1：μ更新前 != μ更新後

* 1. 結論：



P = 0.8560885111177758 > α = 0.05

接受H0 =>

在α = 0.05，更新前後角色勝率無差異的說法是可以支持的。

* 1. 某文化部長想知道109年高雄市各語言分布比例是否與台灣各語言分布有差異？
  2. Data Source：<https://www.thenewslens.com/article/157030>



資料描述：高雄市跟台灣的各語言使用比例

* 1. 分析方法：卡方適合度檢定

為何採用：高雄市的分布為期望值，台灣的分布為觀察值

* 1. 分析結果：

程式碼（Language\_TK\_chi.py）：

from scipy.stats import chisquare

import pandas as pd

df1 = pd.read\_excel('people\_languge.xlsx')

Oi = df1['Oi'].values[0:5]

Ei = df1['Ei'].values[0:5]

#print(Oi, Ei)

print(chisquare(Oi, Ei))

統計假設：

H0：P國語、P台語、P客語、P原住民族語、P其他比例是

66.37：37.73：1.5：0.2：0.2

H1：P國語、P台語、P客語、P原住民族語、P其他比例不是

66.37：37.73：1.5：0.2：0.2

* 1. 結論：



P = 0.0 < α = 0.05

X2 = 166886753.81923237 > X20.05(4) = 9.48773

拒絕H0 =>

在α = 0.05，高雄市各語言分布比例與台灣各語言分布比例無差異的說法是不可以支持的。

* 1. 某文化部長想知道109年台灣各語言使用的比例是否有差異？
  2. Data Source：<https://www.thenewslens.com/article/157030>



資料描述：分為每個語言使用的人數及其使用的語言

語言種類有：國語，台語，客語，原住民語，其他

* 1. 分析方法：卡方適合度檢定

為何採用：X為語言（類別尺度），Y為各種語言使用人數（間斷資料）

* 1. 分析結果：

程式碼（Language\_chi.py）：

from scipy.stats import chisquare

import pandas as pd

df1 = pd.read\_excel('Language.xlsx')

anscount = df1['人數']

print(chisquare(anscount))

統計假設：

H0：P國語 = P台語 = P客語 = P原住民族語 = P其他

H1：P國語 != P台語 != P客語 != P原住民族語 != P其他

* 1. 結論：

P = 0.0 < α = 0.05

拒絕H0 =>

在α = 0.05，109年台灣各語言使用的比例無差異的說法是不可以支持的。

* 1. 高雄市政府想知道各月份的運載量是否有差異？
  2. Data Source：<https://reurl.cc/LppA04>



資料描述： 輕軌民國107年至今每個月份的運載量

* 1. 分析方法：卡方適合度檢定

為何採用：X為運載量（類別尺度），Y為月份（間斷資料）

* 1. 分析結果：

程式碼（train\_chi.py）：

from scipy.stats import chisquare

import pandas as pd

df1 = pd.read\_excel('train.xlsx')

s = df1['總運量']

print(chisquare(s))

統計假設：

H0：各月份運載量分布無差異

H1：各月份運載量分布有差異

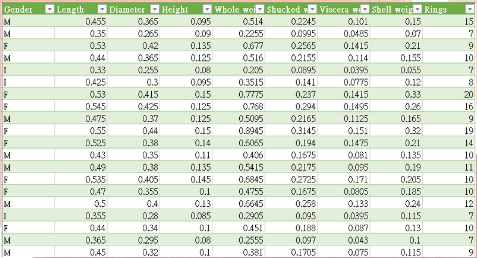
* 1. 結論：

P = 0.0 < α = 0.05

拒絕H0 =>

在α = 0.05，各月份運載量分布無差異的說法是不可以支持的。

* 1. 某海產業者想知道鮑魚在不同年紀時重量是否有差異？
  2. Data Source：<https://archive.ics.uci.edu/ml/datasets/Abalone>



資料描述：年齡斷分為H（> 20），M（19 - 10），L（< 10）

重量指的是一顆鮑魚帶殼的重量單位為公斤

* 1. 分析方法：單因子ANOVA檢定

為何採用：X為三種不同年齡段（類別尺度），Y為鮑魚含殼重量（連續尺度）

* 1. 分析結果：

程式碼（Abalone\_anova.py）：

import scipy.stats as stats

import pandas as pd

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

df1 = pd.read\_excel('Abalone.xlsx')

H = []

M = []

L = []

for i in range(len(df1['Rings'])):

    if df1['Rings'][i] == "H":

        H.append(df1['Whole\_weight'][i])

    elif df1['Rings'][i] == "M":

        M.append(df1['Whole\_weight'][i])

    else:

        L.append(df1['Whole\_weight'][i])

lev = stats.levene(H, M, L)

print('levene test', lev)

if lev[1] > 0.05:

    print('變異數同值')

    ano1 = stats.f\_oneway(H, M, L)

    print(ano1)

    if ano1[1] <= 0.05:

        print('事後比較')

        df2 = df1['Rings']

        # print(df2)

        mst = df1['Whole\_weight']

        tukey = pairwise\_tukeyhsd(endog=mst, groups=df2, alpha=0.05)

        print(tukey.summary())

else:

    krs = stats.kruskal(H, M, L)

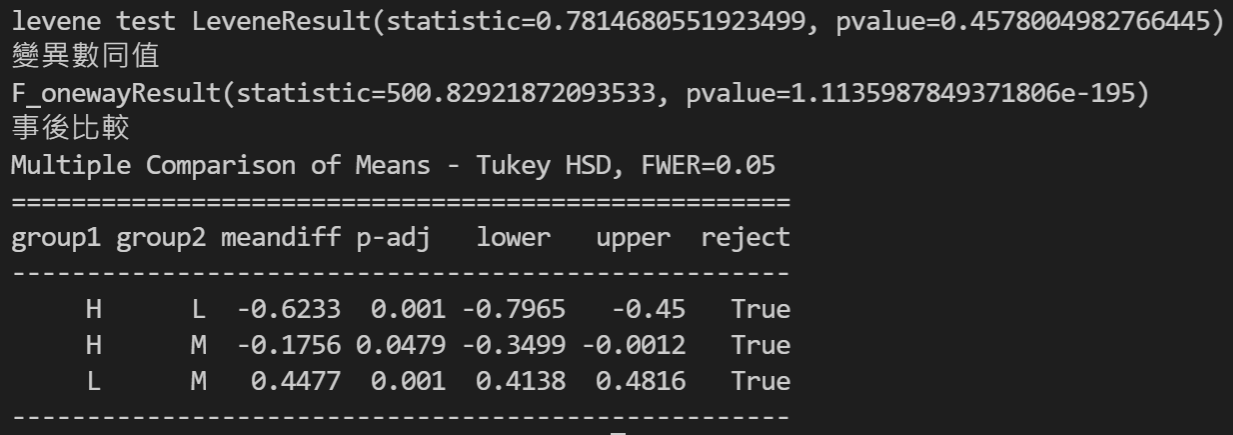
    print('變異數異值')

    print(krs)

統計假設：

H0：μL = μM = μH

H1：μL != μM != μH

* 1. 結論：

Lev：p = 0.4578004982766445 > α = 0.05 => 變異數同質

Oneway：p = 1.1135987849371806e-195 < 0.05 => 拒絕H0

在α = 0.05下，鮑魚在不同年紀時重量無差異的說法是不可以支持的

事後比較：三種年紀各獨立為一組

* 1. 高雄市政府想知道月份對於運載量是否有差異？
  2. Data Source：<https://reurl.cc/LppA04>



資料描述：輕軌民國107年至今為止每個月份的運載量

* 1. 分析方法：單因子ANOVA檢定

為何採用：X為月份（類別尺度），Y為每月總運量（連續尺度）

* 1. 分析結果：

程式碼（train\_onewayanova.py）：

import pandas as pd

import scipy.stats as stats

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

df1 = pd.read\_excel('train.xlsx')

month = df1['月']

s = df1['總運量']

Jan = []

Feb = []

Mar = []

Apr = []

May = []

Jun = []

Jul = []

Aug = []

Sep = []

Oct = []

Nov = []

Dec = []

for i in range(len(df1['月'])):

    if df1['月'][i] == 1:

        Jan.append(df1['總運量'][i])

    elif df1['月'][i] == 2:

        Feb.append(df1['總運量'][i])

    elif df1['月'][i] == 3:

        Mar.append(df1['總運量'][i])

    elif df1['月'][i] == 4:

        Apr.append(df1['總運量'][i])

    elif df1['月'][i] == 5:

        May.append(df1['總運量'][i])

    elif df1['月'][i] == 6:

        Jun.append(df1['總運量'][i])

    elif df1['月'][i] == 7:

        Jul.append(df1['總運量'][i])

    elif df1['月'][i] == 8:

        Aug.append(df1['總運量'][i])

    elif df1['月'][i] == 9:

        Sep.append(df1['總運量'][i])

    elif df1['月'][i] == 10:

        Oct.append(df1['總運量'][i])

    elif df1['月'][i] == 11:

        Nov.append(df1['總運量'][i])

    else:

        Dec.append(df1['總運量'][i])

#print(Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)

lev = stats.levene(Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)

print('levene test', lev)

if lev[1] > 0.05:

    print('變異數同值')

    ano1 = stats.f\_oneway(Jan, Feb, Mar, Apr, May, Jun,

                          Jul, Aug, Sep, Oct, Nov, Dec)

    print(ano1)

    if ano1[1] <= 0.05:

        print('事後比較')

        df2 = df1['月']

        # print(df2)

        mst = df1['總運量']

        tukey = pairwise\_tukeyhsd(endog=mst, groups=df2, alpha=0.05)

        print(tukey.summary())

else:

    krs = stats.kruskal(Jan, Feb, Mar, Apr, May, Jun,

                        Jul, Aug, Sep, Oct, Nov, Dec)

    print('變異數異值')

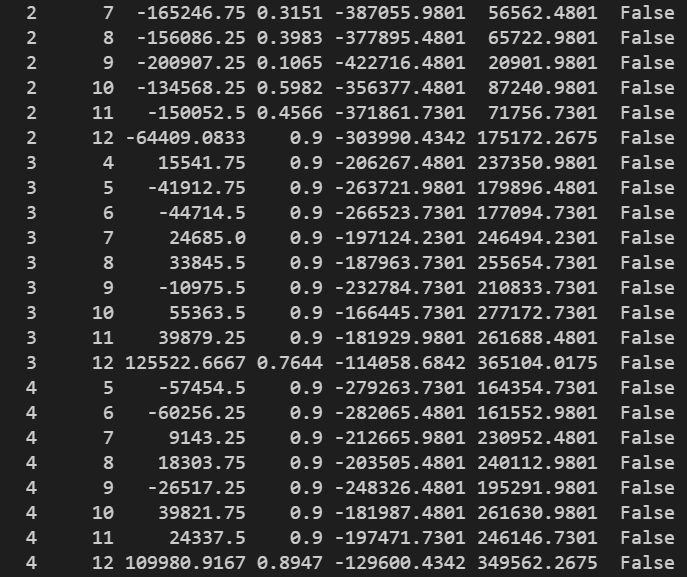
    print(krs)

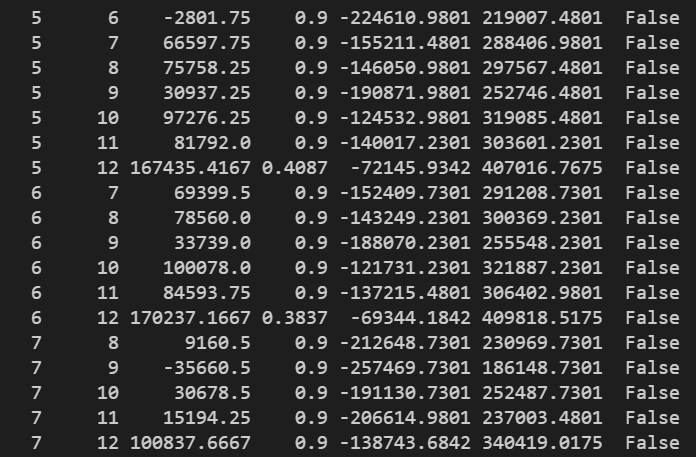
統計假設：

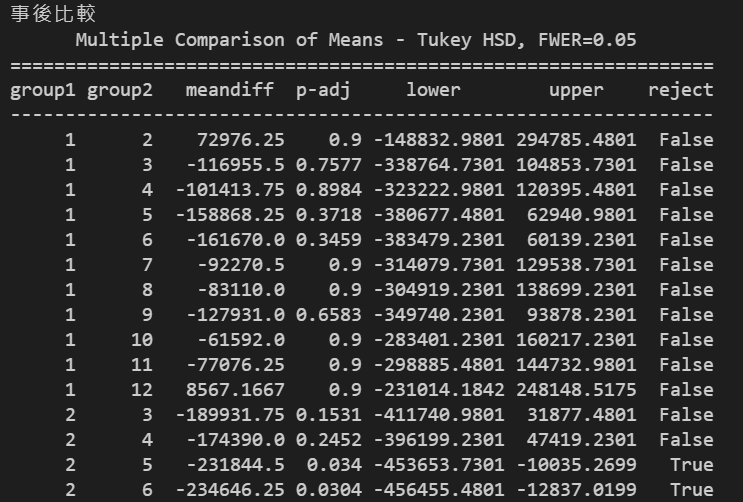
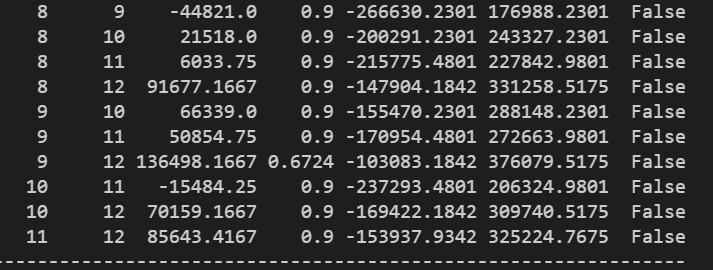
H0：月份和運載量無差異

H1：月份和運載量有差異

* 1. 結論：







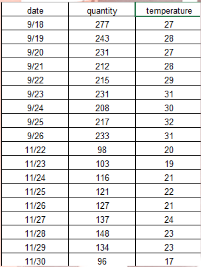
Lev：p = 0.7909309084131676 > α = 0.05 => 變異數同質

Oneway：p = 0.02435079091858752 < 0.05 => 拒絕H0

在α = 0.05下，輕軌在不同月份時總運量無差異的說法是不可以支持的。

事後比較：2月與5月、2月與6月，非為一組；其餘的組合則反之

* 1. 我們欲檢視飲料杯數是否隨著氣溫增加而遞增，即在顯著水準0.05下，檢定斜率β1是否等於零？
  2. Data Source：<https://reurl.cc/bkkGRr>



資料描述：冬季、夏季的氣溫&飲料銷售杯數

* 1. 分析方法：回歸分析

為何採用：X為氣溫（連續尺度），Y為飲料銷售杯數（連續尺度）

* 1. 分析結果：

程式碼（Drink\_R.py）：

import numpy as np

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

import pandas as pd

df = pd.read\_excel('drink.xlsx')

data\_x = df['temperature']

data\_y = df['quantity']

# 分割trian/test

train\_x, test\_x, train\_y, test\_y = train\_test\_split(data\_x,

                                                    data\_y,

                                                    test\_size=0.25,

                                                    random\_state=1)

# 氣溫對杯數的散佈圖

plt.scatter(train\_x, train\_y, facecolor='None', edgecolor='k', alpha=0.3)

plt.show()

train\_x.values.reshape(-1, 1)

train\_x\_reshape = train\_x.values.reshape(-1, 1)

model = LinearRegression()

model.fit(train\_x\_reshape, train\_y)

print(f"Coefficient:{model.coef\_}")

print(f"Beta0:{model.intercept\_}")

print(f"R2:{model.score(train\_x\_reshape, train\_y)}")

data\_x\_reshape = data\_x.values.reshape(-1, 1)

xp = np.linspace(data\_x\_reshape.min(), data\_x\_reshape.max(), 70)

xp = xp.reshape(-1, 1)

pred\_plot = model.predict(xp)

plt.title('LinearRegression')

plt.xlabel('Temperature')

plt.ylabel('Quantity')

plt.scatter(data\_x\_reshape, data\_y, facecolor='None', edgecolor='k', alpha=0.8)

plt.plot(xp, pred\_plot)

plt.show()

test\_x\_reshape = test\_x.values.reshape(-1, 1)

pred = model.predict(test\_x\_reshape)

print("predict")

for i in range(5):

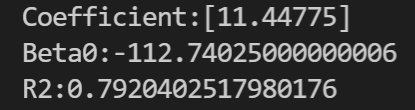
    print("true:", test\_y.values[i])

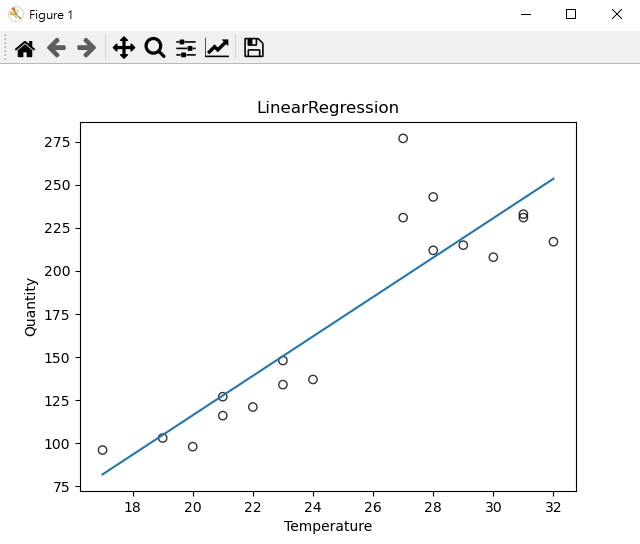
    print("prediction:", pred[i])

統計假設：

H0：氣溫與飲料杯數不呈現性關係

H1：氣溫與飲料杯數呈現性關係

* 1. 結論：



Ŷ=-112.74025+11.44775X

R-squard：0.7920402517980176，表示此模型在X->Y的解釋能力為79.2%

Coefficient：11.44775 > 0 => 正相關

一張含有 文字 的圖片

自動產生的描述 預測：

氣溫對飲料店的銷售飲料杯數呈正線性關係